**Title**

Generating Polyphonic Music with Attention-Based Transformer Model

**Topic Area**

Many deep learning models for music generation are based on recurrent neural networks (RNNs), which can struggle to capture long-term dependencies in music. This thesis aims to explore the potential of attention-based transformer models for generating polyphonic music, which can capture more complex relationships between musical elements over longer time periods.

The field of music generation has witnessed remarkable advancements in recent years, largely driven by the application of deep learning techniques. One such approach is the attention-based Transformer model, originally introduced in the "Attention Is All You Need" paper (Vaswani *et al.*, 2017). This model has shown great promise in natural language processing tasks and has been successfully adapted for music generation.

The Transformer model, originally designed for natural language processing, has demonstrated its ability to capture long-range dependencies and understand the hierarchical structure of sequences. This makes it a promising candidate for modeling the complex relationships between multiple voices in polyphonic music. By leveraging the self-attention mechanism, the Transformer can attend to relevant parts of the input sequence and generate coherent and harmonically-rich music compositions.

The application of attention-based Transformer models to polyphonic music generation opens up exciting opportunities for creating original and expressive compositions. Researchers and practitioners in the field are actively exploring novel approaches and techniques to further enhance the capabilities of these models. By studying and advancing the generation of polyphonic music with attention-based Transformer models, we can unlock new avenues for creative musical expression and contribute to the development of intelligent systems that can compose complex and engaging music.

**Research Objectives**

* Problem Identification: Transformer models have shown great promise in generating polyphonic music, but it is unclear how well they can capture long-term dependencies in the music.

Problem Clarification: The ability to capture long-term dependencies is important in generating music that has a coherent structure and is musically pleasing.

Problem Formulation: How effective are self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music?

Objective: To evaluate the effectiveness of self-attention mechanisms in transformer models for capturing long-term dependencies in polyphonic music.

* This objective aims to investigate the suitability of transformer models for music generation by evaluating their ability to capture long-term dependencies in polyphonic music. The objective will involve examining the effectiveness of self-attention mechanisms in identifying and encoding relationships between musical elements over longer time periods.
* Problem Identification: The impact of training data on the performance of attention-based transformer models for polyphonic music generation is not well understood.

Problem Clarification: It is unclear how the size and diversity of training data affect the quality of generated music by attention-based transformer models.

Problem Formulation: The objective is to evaluate the impact of training data on the performance of attention-based transformer models for polyphonic music generation, and how the size and diversity of training data affects the quality of generated music.

Objective: To determine the relationship between training data size and diversity, and the quality of generated music by attention-based transformer models for polyphonic music generation.

* The objective of this study is to investigate how the quality of generated music by attention-based transformer models for polyphonic music generation is affected by the size and diversity of the training data used to train these models. In other words, the study aims to determine whether the quantity and variety of the training data have an impact on the quality of the generated music. By evaluating this relationship, the study can provide insights into how to optimize the training data selection process to improve the performance of attention-based transformer models for polyphonic music generation.
* Problem Identification: Computer-generated music often suffers from repetitive patterns, which can make the music uninteresting and predictable.

Problem Clarification: Generating diverse and original music is important in creating music that is musically pleasing and engaging.

Problem Formulation: How well can attention-based transformer models generate diverse and original polyphonic music, and how effective are they at avoiding repetitive patterns and generating novel musical ideas?

Objective: To assess the ability of attention-based transformer models to generate diverse and original polyphonic music by examining their ability to avoid repetitive patterns and generate novel musical ideas.

* The objective of this research is to evaluate the diversity and originality of polyphonic music generated by attention-based transformer models. Specifically, the research aims to examine the model's ability to avoid repetitive patterns and generate novel musical ideas. By assessing the model's ability to generate diverse and original music, this research can contribute to the development of more advanced and creative machine learning models for music generation.

**Literature Review**

**The Literature Review is incomplete and heads have been placed as placeholders for sections to be added. The section that has been complete is self-attention transformer.**

Choosing a model

* Given a music sequence x = [x1, x2, ..., xn], we choose to model the unknown data distribution pdata(x) autoregressively, i.e., pθ(x) = ∏n t=1 pθ(xt|x1, ..., xt−1). The most common approach to train such a model is Maximum Likelihood that finds a θ to minimize, Lmle = Ex∼pdata(x) − log(pθ(x)). (1) Despite its attractive theoretical properties, Maximum Likelihood training suffers from many limitations, e.g. whenever the model is misspecified. This issue is illustrated by (Isola et al., 2017) (Muhamed *et al.*, 2021)

RNN

* LSTM-based model like Performance RNN that compresses earlier events into a fixed-size hidden state.
* The model seems to “forget” about the primer almost immediately. While LSTM-based models are able to generate music that sounds plausible at time scales of a few seconds or so, the lack of long-term structure is apparent. As a consequence, Performance RNN is unable to generate coherent continuations to a user-specified primer performance. (*Music Transformer: Generating Music with Long-Term Structure*, 2018)
* [folk-rnn](https://github.com/IraKorshunova/folk-rnn) models(Sturm, 2017)
* Pianoroll RNN-NADE
* Neural Autoregressive Distribution Estimation (NADE) is a specific type of model that falls under the broader category of autoregressive models. Autoregressive models generate data one step at a time, where each step depends on the previous steps. In the context of music generation, NADE uses neural networks to estimate the probability distribution of the next note or event in a sequence given the preceding notes. By training on a large dataset of music, NADE can learn to generate new music that shares similarities with the training data. (chat gpt)
* This model applies language modeling to polyphonic music generation using an LSTM combined with a NADE, an architecture called an RNN-NADE. Unlike melody RNNs, this model needs to be capable of modeling multiple simultaneous notes. It does so by representing a NoteSequence as a "pianoroll", named after the medium used to store scores for player pianos.
* In a pianoroll, the score is represented as a binary matrix where each row represents a step and each column represents a pitch. When a column has a value of 1, that means the associated pitch is active at that time. When the column value is 0, the pitch is inactivate. A downside of this representation is that it is difficult to represent repeated legatto notes since they will appear as a single note in a pianoroll.
* Since we need to output multiple pitches at each step, we cannot use a softmax. The polyphony\_rnn posted previously skirted this issue by representing a single time step as multiple, sequential outputs from the RNN. In this model, we instead use a Neural Autoregressive Distribution Estimator, or NADE to sample multiple outputs given the RNN state at each step. See the original RNN-NADE paper and our code for more details on how this architecture works.(*magenta/magenta/models/pianoroll\_rnn\_nade at main · magenta/magenta*, no date) (Sigtia, Benetos and Dixon, 2016)
* char-rnn (Sturm, 2017)
* However, it has been a challenge to equip neural networks with the capability to model long-term dependency in sequential data. Recurrent neural networks (RNNs), in particular Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have been a standard solution to language modeling and obtained strong results on multiple benchmarks. Despite the wide adaption, RNNs are difficult to optimize due to gradient vanishing and explosion (Hochreiter et al., 2001), and the introduction of gating in LSTMs and the gradient clipping technique (Graves, 2013) might not be sufficient to fully address this issue. Empirically, previous work has found that LSTM language models use 200 context words on average (Khandelwal et al., 2018), indicating room for further improvement.(Dai *et al.*, 2019)

SELF-ATTENTION TRANSFORMER

Attention-based transformer models have been proposed as an alternative to RNNs for capturing long-term dependencies in music. Transformers are a type of deep learning model which has been a powerful tool for various natural language processing tasks, including language translation and text generation. (Vaswani et al., 2017) The Transformer model is a new kind of encoder-decoder model that uses self-attention to make sense of language sequences. This allows for parallel processing and thus makes it much faster than any other model with the same performance. They thus paved the way for modern language models (such as BERT (Devlin *et al.*, 2019), GPT (Brown *et al.*, 2020) and T5 (Raffel *et al.*, 2020)) Attention-based transformer models are based on the self-attention mechanism, which allows the model to attend to different parts of the input sequence when making predictions. The self-attention mechanism in attention-based transformer models enables the model to weigh the importance of different positions in the input sequence, allowing it to focus on the most relevant information for a given task. This makes the model more flexible and adaptable to different input lengths and patterns compared to RNNs. This mechanism allows the model to capture long-term dependencies without the vanishing gradient problem seen in RNN models. However, this model is not very practical to implement for music generation because of how computational expense it is to run. i.e. the square of the sequence length (Huang *et al.*, 2018) proposed an algorithm that reduces their intermediate memory requirement to linear in the sequence length. This enabled them to demonstrate that a Transformer with modified relative attention mechanism can generate minute long compositions with compelling structure, generated continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies. However, even though transformers have a potential of learning longer-term dependency, they are also limited by a fixed-length context in the setting of language modelling. (Dai *et al.*, 2019) proposed a novel neural architecture Transformer-XL that enables learning dependency beyond a fixed length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. Their method not only enables capturing longer-term dependency, but also resolves the context fragmentation problem. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers A.K.A the original transformer by (Vaswani *et al.*, 2017), achieves better performance on both short and long sequences, and is up to 1,800+ times faster than vanilla Transformers during evaluation. Leveraging off of this discovery (Donahue *et al.*, 2019) used transfer learning procedure to generate video game sound synthesis chip music which is multi-instrumental music by pre-training on a widely used large-scale dataset called Lakh MIDI and using Transformer-XL. Showing that they improve results both quantitatively and qualitatively by pertaining on a cross-domain dataset. As well as generating both chiptunes from scratch and collaborating with human composers. (Wu, Wang and Lei, 2020) built on this idea of Transformer-XL and instead of a single sequence-based Transformer-XL they generated music with multiple sequence of time-valued notes. They aimed to address two challenges: computing notes with the same value but different tempos, and the model's limitation in separately learning music aspects like harmony and rhythm due to the use of a single sequence. However, by solving these issues it is adding more complexity to the transformer which can lead to longer training times and higher computational requirements. Similarly, (Ens and Pasquier, 2020) is a model that explores conditional multi-track music generation using the Transformer architecture. Although it is limited by its ability to only generate music for a fixed number of bars it has created a space to explore diverse musical styles and the creation of complex and engaging multi-track music. (*MuseNet*, 2019) developed by OpenAI, is a prominent AI model that employs a variant of the Transformer architecture to generate diverse and original musical compositions. Trained on a vast dataset of MIDI files, MuseNet demonstrates the ability to produce coherent and stylistically diverse compositions in various genres, offering potential applications in music production, creative inspiration, and entertainment. It only drawback is that it lacks a deep understanding of musical theory and context, sometimes resulting in compositions that may sound musically inconsistent or unfamiliar. Contrastingly, (Huang and Yang, 2020) aim to focus on just Pop piano music and build a Pop Music Transformer. The paper highlights the effectiveness of the Transformer model for generating expressive classical piano performances, while proposing improvements in data representation to enhance music modeling. By incorporating a metrical structure in the input data, the Pop Music Transformer is developed to generate Pop piano music with improved rhythmic and harmonic structures. These advancements showcase the Transformer's ability to learn abstractions and generate coherent compositions without relying heavily on human-imposed constraints or domain knowledge. The only issue with this approach is that it makes it less versatile for generating music outside the pop genre. Likewise, the work done on music generation becomes null and void if it is not interacted with by musicians. Generative algorithms are still not widely used by artists due to the limited control they offer, prohibitive inference times or the lack of integration within musicians’ workflows.(Hadjeres and Crestel, 2021) tackles this by presenting the Piano Inpainting Application (PIA), a generative model focused on “inpainting” piano performances, as they believe that elementary operation (restoring missing parts of a piano performance) encourages human-machine interaction and opens up new ways to approach music composition. It allows musicians to smoothly generate or modify any MIDI clip using PIA within a widely used professional Digital Audio Workstation. The transformer used in this approach is the Linear Transformer by (Katharopoulos *et al.*, 2020) which is performs similar to the original vanilla Transformer but can be up to 4000x faster on autoregressive prediction of very long sequences. (Choi *et al.*, 2020) explores the use of Transformer autoencoders to encode and decode musical style representations. By leveraging the Transformer architecture, the model can learn to capture and reconstruct the underlying stylistic elements of music, allowing for style manipulation and generation. The approach shows promise in enabling fine-grained control over musical style and opens up possibilities for style transfer and composition. As well as music being generated in a specific style a crucial element of make great music is when it evokes an emotion from the listener. (Makris, Agres and Herremans, 2021) introduces a novel approach to generating lead sheets that incorporates high-level musical characteristics, specifically valence (the positivity or negativity of the perceived emotion), for control over the generated output. By using pre-defined mood tags and a conditional sequence-to-sequence framework, the authors demonstrate the ability to generate lead sheets in a controllable manner, achieving distributions of musical attributes similar to the training data and effective control over the valence of the generated chord progressions. This human-like element to the music being generated brings us closer to creating emotionally expressive and authentic compositions that resonate with listeners on a deeper level, bridging the gap between machine-generated music and the artistry of human musicians. Finally, the model that appears to yield the most favorable results in comparison to the Vanilla Transformer and Music Transformer is the Transformer-GANs. Drawing on a diverse range of methodologies explored above, it emerges as a compelling solution, offering superior outcomes in terms of music generation by leveraging the strengths of both the Transformer architecture and Generative Adversarial Networks. (Muhamed *et al.*, 2021), (Muhamed *et al.*, 2021) (these are different papers with the same name and same main author) and (Neves, Fornari and Florindo, 2022) use this approach with great success. They propose a framework that combines the Transformer architecture and Generative Adversarial Networks (GANs) for generating music with specific sentiments. The authors introduce a sentiment encoder to condition the generation process, allowing control over the emotional content of the generated music. Experimental results demonstrate the effectiveness of their approach in producing music that conveys desired sentiments while maintaining musical coherence and quality. However, a huge drawback and reason why we would not use it in our strategy is Transformer-GANs is very complex and computational costly when training the model. Transformer-GANs combine the Transformer architecture with Generative Adversarial Networks, which both require significant computational resources and training time. This can pose challenges in terms of scalability and practicality, especially when dealing with large datasets or real-time music generation scenarios. Additionally, ensuring the stability and convergence of the GAN training process can be challenging, requiring careful tuning and experimentation.

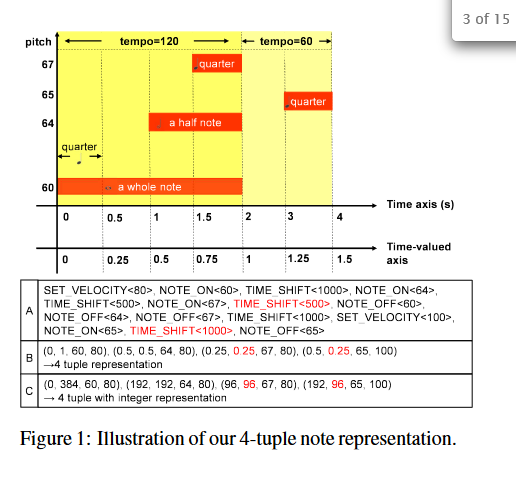
Finally, we compare the evaluation speed of our model with the vanilla Transformer model (AlRfou et al., 2018). As shown in Table 9, due to the state reuse scheme, Transformer-XL achieves an up to 1,874 times speedup during evaluation.(Dai *et al.*, 2019)

* Theme transformer: A music transformer will prompt conditioned music generation

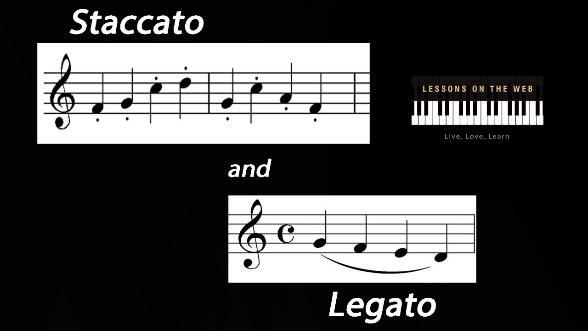
DATASET

* Symbolic music
* Audio music
* sequential data.
* Audio files
* Midi files
* Quatise midi files
* Human midi files
* Polyphonic music
* Predicting chords, is it the whole chord, top note, middle note, last note?
* Important considerations for data cleaning and processing include handling **outliers**, **dealing with variable sequence lengths**, and **encoding the data** in a form that can be processed by the model. **Time signatures**, **key signatures**, and **tempo** may also need to be encoded, depending on the complexity of the music and the specific goals of the project.

MUSIC



* Monoponic, homophonic, polyphonic, heterophonic
* Expressive **timing** and **dynamics** are an essential part of music. Listen to the following two clips of the same Chopin piece, the first of which has been stripped of these qualities:
* The first clip is just a direct rendering of the score, but with all notes at the same volume and quantized to 16th notes. The second clip is a MIDI-recorded human performance with phrasing. Notice how the same notes lead to an entirely different musical experience. That difference motivates this work.
* Performance RNN generates expressive timing and dynamics via a stream of MIDI events. At a basic level, MIDI consists of precisely-timed note-on and note-off events, each of which specifies the pitch of the note. Note-on events also include velocity, or how hard to strike the note.
* These events are then imported into a standard synthesizer to create the “sound” of the piano. In other words, the model only determines which notes to play, when to play them, and how hard to strike each note. It doesn’t create the audio directly.(*Performance RNN: Generating Music with Expressive Timing and Dynamics*, 2017)



* *why is it in music generation that in a 4 4 bar that the durstion is commonly 16*
* In a 4/4 time signature, the bar is divided into four beats, and each beat is typically represented by a quarter note. If you further divide each beat into four equal parts, you get sixteenth notes. So, in a bar of 4/4 time, there are a total of 16 sixteenth notes.

This subdivision into 16 parts provides flexibility and allows for more intricate rhythms and patterns. By using sixteenth notes, musicians can add complexity to the music, and it gives composers a lot of freedom in terms of the rhythmic patterns they can create. It's also useful for aligning various musical elements and creating syncopation and other interesting rhythmic effects.

This doesn't mean that every bar in 4/4 time must have 16 sixteenth notes; rather, it is just a common way to represent the underlying rhythmic structure. Different musical genres and styles might use this subdivision in various ways.

REPETITIVE PATTERN

* Motif and theme transformers
* Sampling
* Techniques such as temperature scaling during **sampling**, **top-k sampling**, or **nucleus sampling** can help.
* **Regularization techniques** like dropout or **weight decay** can also be used to prevent **overfitting** and encourage the model to explore different parts of the musical space.
* Temperature - Can we control the output of the model at all? Generally, this is an open research question; however, one typical knob available in such models is a parameter referred to as temperature that affects the randomness of the samples. A temperature of 1.0 uses the model’s predicted event distribution as is. This is the setting used for all previous examples in this post.(*Performance RNN: Generating Music with Expressive Timing and Dynamics*, 2017)
* Compared with image generation, both of them emphasize “true”, while music generation emphasizes “quality” rather than “quantity”, and “new” rather than “old”. “New” requires the model to be creative instead of repeating the learned segments all the time. The existing evaluation of “quality” is very subjective. In order to ensure the novelty of the generated melody, we evaluate these melodis in terms of mathematical statistics and music theory knowledge [[28](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283103#pone.0283103.ref028)]. The evaluation model-MEM (Music evaluation model) propoesed in this paper is divided into the following mathematical statistics test and music theory evaluation.(Guo *et al.*, 2023)

How do we represent music? One note? A chord? Is silence important?

Chords

* Chords, as the name might suggest, are objects that combine multiple [**Pitch**](https://web.mit.edu/music21/doc/moduleReference/modulePitch.html#music21.pitch.Pitch) objects on a single stem.
* But since a Chord contains many pitches, it does not have a .pitch attribute:
* Instead it has a [**.pitches**](https://web.mit.edu/music21/doc/moduleReference/moduleChord.html#music21.chord.Chord.pitches) attribute which returns a Tuple of pitches in the Chord.
* (*User’s Guide, Chapter 7: Chords — music21 Documentation*, no date)

**Sampling Strategy**

In this study on the use of attention-based transformer model for polyphonic music generation, our primary research population will consist of five industry experts. An industry expert is defined as someone who has a least completed their Bachelor of Arts degree in music or worked in the music industry for 10 years. We are sampling industry experts because we want to gain valuable insights from their in-depth knowledge of the creation of music that can provide a nuanced perspective to enrich our research. They offer a credible and reliable point of view that we can trust. Finally, they provide a unique perspective on the music generated that we might not notice for ourselves.

These experts will be gathered from Linked In and will not be a connection on Linked In prior to the research to avoid any response bias. Linked In is used as the sample frame of choice because it provides a large rich pool of experts due it’s global popularity and widespread adoption among professionals from various industries and geographical locations. It is easily accessible and owing to its user-friendly interface, it is virtually effortless to select a sample.

Due to the fact that this research is under time constraints we will be selecting the non-probability sampling technique and the sampling type will be judgement sampling more specifically expert sampling. Although the probability sampling method would provide a representative sample and allow for statistical generalization. We choose non-probability sampling instead because it is a very efficient method of retrieving feedback needed in a short space of time. The reason for selecting judgement sampling is because we are conducting exploratory research where we want to hand pick the people, we feel have the right collection of experience and proficiency to give us worthwhile feedback on the music we generated. The experts will be selected in such a way that to ensure diversity in expertise, perspectives, and backgrounds i.e., obtain different job titles in various companies. The research will be relevant to our overall research as the experts will give their opinion on whether the music is coherent, the dataset affects the quality of the music, or the music is monotonously repetitive. This will tie our research back to our objectives for our overall research. The experts will have a well-tuned ear to the subtle nuances and intricacies of the music, allowing them to provide insightful and knowledgeable input, which ultimately saves us time and resources.

**Primary Research Methodology**

The method used in our primary research is five expert interviews. The experts will be reached out to via Linked In and invited to participate in the study. We will share the topic of interest, research objectives and music generated from our analysis for the interviewees to listen to prior to the interview. The interview will take approximately forty-five minutes in total and will allow fifteen minutes for use to summarise our initial thoughts after the interview is over. Consent for the recording of the interview will be received prior to the interview taking place via email/ Linked In message.

The expert interviews will be conducted using a qualitative approach. Depending on logistical concerns, the interviews will be organized for a mutually convenient time and may be performed in person in a quiet public setting e.g., cafe, over the phone, or via zoom. The experts will have a chance to share their knowledge and opinions during the interviews on both the music we have generated and whether it meets our objectives and music generation in general, to give us a better understand on the topic. To ensure the information collected from the interview is relevant to our research. The open-ended questions will be designed to elicit detailed and insightful responses from the experts, allowing for exploration and clarification of their perspectives.

The interviews' audio recordings will be transcribed verbatim, and the transcripts will be used as the main source of analysis data. To detect patterns, themes, and significant results in the interview data, thematic analysis, content analysis, or other qualitative analytic techniques will be used. The reliability and validity of the results will be guaranteed by the application of rigorous analysis procedures. When the analysis is complete it will be compared to the secondary research finding to see do our results match what the experts have found and if not why.

The reason for our preferred method of primary research is because we need more than just someone to tell us if the music we have generated is subjectively good or not. To make informed decisions and draw meaningful conclusions from our research findings, we require someone with a valid and relevant opinion, whose evaluation holds credibility. We especially need this in our first objective where is see if our self-attention transformer can capture long-term dependencies in polyphonic music. The expert will evaluate whether the newly generated music is a coherent piece that seamlessly integrates with the trained music. They will determine if it is a well-structured composition or merely a sequence of random notes. We need an expert to pinpoint aspects of the music that are done well or more importantly where it has not. Particularly for our second objective for whether the music changes in quality when a new dataset has been used on the model. The expert is key here as they know what makes up good quality music and would be able to hear instantly whether there is a significant change. Certain metrics might be considered such as pitch, rhythm, melody, harmony, originality, and emotional impact. Finally, conducting an expert interview provides flexibility with the questions we ask. If the interviewee brings up something that captures our interest and hasn't been previously considered, we can explore it further. Similarly, if there are aspects that we find unclear, we can ask the interviewee to elaborate on the spot. This specific advantage of the research method will prove helpful when discussing the third objective as there is a fine line between music repeating itself for structure, emphasis, rhythm and so on without it being monotonous and boring to the listener. Here the expert can delve into this deeper with us and explain whether the balance has been met.

While employing this research method, it is essential to acknowledge certain limitations. Firstly, the subjective nature of expert opinions introduces the possibility of biases, as feedback is influenced by their individual experiences, knowledge, and personal inclinations. Additionally, the findings may lack generalizability to the broader population, as we have only interviewed a limited sample of five experts rather than a larger or more representative group. Nevertheless, despite these limitations, the benefits offered by this method outweigh the drawbacks, leading us to proceed with its implementation in our research.

**Ethics**

There are some ethical considerations we expert to encounter in the undertaking of the Data Analysis project. For the primary research we will need voluntary participation and informed consent from all five of the experts that will be participating in the interviews. We will do this by clearly explaining the purpose of the study, the nature of their participation, and how their data will be used through email/Linked In message. We will encourage them that if they have any questions before hand, they can ask through the same email chain and inform them that we respect their decision to participate or withdraw from the study at any point.

Trust and respect are at the utmost importance for all participants in the study. This means showing up to the meeting early and prepared to start with plenty of questions to fill the time slot agreed. Maintaining professionalism throughout the interview and ensuring their opinions and expertise are valued.

Ensuring the accuracy of result reporting is paramount. This process entails transcribing the recordings verbatim and extracting vital themes, quotes, and interesting insights while avoiding the inclusion of redundant or fabricated information.

When writing up the thesis results, we will preserve the confidentiality and anonymity of the experts when reporting their responses by allocating them an alias i.e., interviewee A. All other information about the interviewee, name, email address, Linked In profile, phone number, etc will be stored in our own password protected private Linked In and/or email accounts and will be stored in a password protected folder on our personal laptop. Lastly, we will share the findings in a manner that respects the experts' contributions while maintaining the integrity of the research. We will report on the themes and insights that were in all five interviews in a summary like format rather than a line-by-line transcript. This will be easier and more interesting to read the key points made in all the interviews while also staying true to what was said by each interviewee.

In conclusion, the Data Analysis project necessitates careful consideration of ethical principles. We will prioritize voluntary participation and informed consent, ensuring that all five experts involved in the interviews fully understand the study's purpose and their role, as well as how their data will be utilized. Trust and respect will be central, demonstrated through punctuality, preparedness, and a genuine appreciation for the experts' opinions and expertise. Reporting and dissemination will adhere to responsible practices, accurately conveying the results of the data analysis by transcribing recordings verbatim and extracting key themes, quotes, and insights. To protect confidentiality and anonymity, aliases will be assigned to experts, while personal information will be securely stored. Ultimately, the findings will be shared in a format that honours the experts' contributions, presenting summarized themes and insights that maintain the integrity of the research and engage readers effectively. By adhering to these ethical considerations, ensures a robust and respectful execution of the Data Analysis project.

**References**

Brown, T.B. *et al.* (2020) ‘Language Models are Few-Shot Learners’. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14165.

Choi, K. *et al.* (2020) ‘Encoding Musical Style with Transformer Autoencoders’. arXiv. Available at: https://doi.org/10.48550/arXiv.1912.05537.

Dai, Z. *et al.* (2019) ‘Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context’. arXiv. Available at: https://doi.org/10.48550/arXiv.1901.02860.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.

Donahue, C. *et al.* (2019) ‘LakhNES: Improving multi-instrumental music generation with cross-domain pre-training’. arXiv. Available at: https://doi.org/10.48550/arXiv.1907.04868.

Ens, J. and Pasquier, P. (2020) ‘MMM : Exploring Conditional Multi-Track Music Generation with the Transformer’. arXiv. Available at: https://doi.org/10.48550/arXiv.2008.06048.

Guo, Y. *et al.* (2023) ‘An automatic music generation and evaluation method based on transfer learning’, *PLOS ONE*, 18(5), p. e0283103. Available at: https://doi.org/10.1371/journal.pone.0283103.

Hadjeres, G. and Crestel, L. (2021) ‘The Piano Inpainting Application’. arXiv. Available at: https://doi.org/10.48550/arXiv.2107.05944.

Huang, C.-Z.A. *et al.* (2018) ‘Music Transformer’. arXiv. Available at: http://arxiv.org/abs/1809.04281 (Accessed: 31 March 2023).

Huang, Y.-S. and Yang, Y.-H. (2020) ‘Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions’. arXiv. Available at: http://arxiv.org/abs/2002.00212 (Accessed: 15 May 2023).

Katharopoulos, A. *et al.* (2020) ‘Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention’. arXiv. Available at: https://doi.org/10.48550/arXiv.2006.16236.

*magenta/magenta/models/pianoroll\_rnn\_nade at main · magenta/magenta* (no date) *GitHub*. Available at: https://github.com/magenta/magenta/tree/main/magenta/models/pianoroll\_rnn\_nade (Accessed: 20 July 2023).

Makris, D., Agres, K.R. and Herremans, D. (2021) ‘Generating Lead Sheets with Affect: A Novel Conditional seq2seq Framework’. arXiv. Available at: https://doi.org/10.48550/arXiv.2104.13056.

Muhamed, A. *et al.* (2021) ‘Symbolic Music Generation with Transformer-GANs’, *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1), pp. 408–417. Available at: https://doi.org/10.1609/aaai.v35i1.16117.

*MuseNet* (2019). Available at: https://openai.com/research/musenet (Accessed: 15 May 2023).

*Music Transformer: Generating Music with Long-Term Structure* (2018) *Magenta*. Available at: https://magenta.tensorflow.org/music-transformer (Accessed: 16 July 2023).

Neves, P., Fornari, J. and Florindo, J. (2022) ‘Generating music with sentiment using Transformer-GANs’. arXiv. Available at: http://arxiv.org/abs/2212.11134 (Accessed: 14 May 2023).

*Performance RNN: Generating Music with Expressive Timing and Dynamics* (2017) *Magenta*. Available at: https://magenta.tensorflow.org/performance-rnn (Accessed: 20 July 2023).

Raffel, C. *et al.* (2020) ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’. arXiv. Available at: https://doi.org/10.48550/arXiv.1910.10683.

Sigtia, S., Benetos, E. and Dixon, S. (2016) ‘An End-to-End Neural Network for Polyphonic Piano Music Transcription’. arXiv. Available at: http://arxiv.org/abs/1508.01774 (Accessed: 20 July 2023).

Sturm, B.L.T. (2017) ‘Even more Endless Music Sessions’, *Folk the Algorithms*, 14 June. Available at: https://highnoongmt.wordpress.com/2017/06/14/even-more-endless-music-sessions/ (Accessed: 20 July 2023).

*User’s Guide, Chapter 7: Chords — music21 Documentation* (no date). Available at: https://web.mit.edu/music21/doc/usersGuide/usersGuide\_07\_chords.html (Accessed: 15 August 2023).

Vaswani, A. *et al.* (2017) ‘Attention Is All You Need’. arXiv. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 1 April 2023).

Wu, X., Wang, C. and Lei, Q. (2020) ‘Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes’. arXiv. Available at: https://doi.org/10.48550/arXiv.2007.07244.